RBI

Santhosh Kumar Krishnan

5/21/2020

1.READ DATA AND INSTALL LIBRARIES

```
RBI <- read.csv("D:/Data Science/Dr Vinod online classes/Class 5-Cluster/RBI.csv")
head(RBI,3)
```

	States_Union <fctr>	BirthRate <dbl>	MortalityRate <int>	PowerAvailability <dbl>	Roa
1	Andaman & Nicobar Islands	12.0	20	473.8	
2	Andhra Pradesh	16.8	37	1019.8	
3	Arunachal Pradesh	18.8	30	427.5	

3 rows

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3
WBa
```

```
library(cluster)
```

2.REMOVE THE RESPONSE VARIABLE

```
df <- RBI[,-1]
rownames(df) <- RBI[,1]
head(df,3)
```

	BirthRate <dbl>	MortalitityRate <int>	PowerAvailability <dbl>	Road
Andaman & Nicobar Islands	12.0	20	473.8	
Andhra Pradesh	16.8	37	1019.8	
Arunachal Pradesh	18.8	30	427.5	

3 rows

3.CHECK NAs AND SCALE THE DATA

```
summary(df)#no NAs
```

```
##    BirthRate     MortalitityRate PowerAvailability   RoadLength
##  Min.   :12.00   Min.   : 9.00   Min.   :  227.9   Min.   :   214
##  1st Qu.:15.10   1st Qu.:19.75   1st Qu.:  471.9   1st Qu.: 14311
##  Median :16.90   Median :27.00   Median :  985.2   Median : 50940
##  Mean   :18.27   Mean   :27.92   Mean   : 1690.6   Mean   :127004
##  3rd Qu.:21.18   3rd Qu.:36.25   3rd Qu.: 1442.5   3rd Qu.:216547
##  Max.   :26.70   Max.   :50.00   Max.   :17281.5   Max.   :608140
```

```
df <- scale(df)#scaled data
head(df)
```

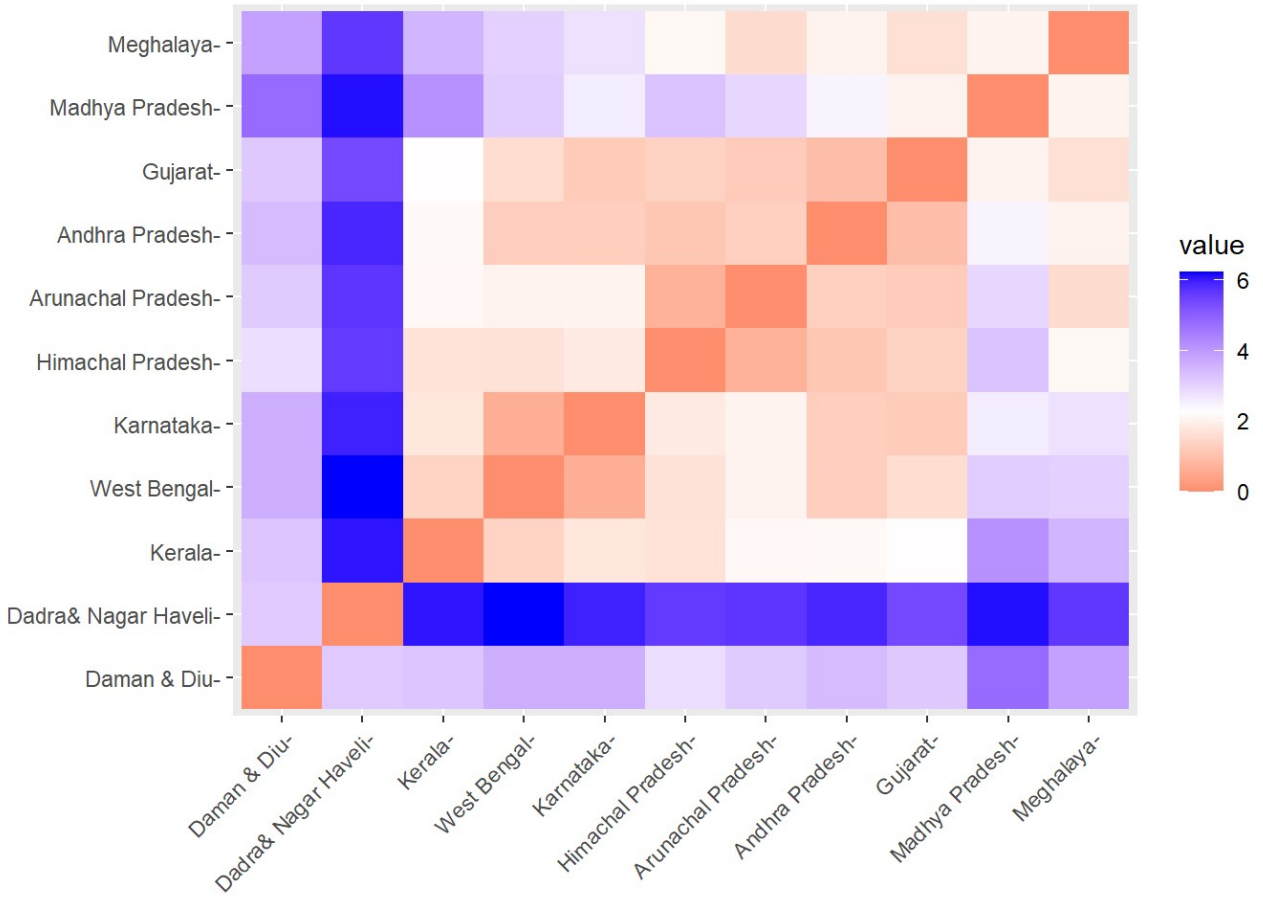
```
##                            BirthRate MortalityRate PowerAvailability
## Andaman & Nicobar Islands -1.4924049     -0.6737197       -0.39161250
## Andhra Pradesh            -0.3502988      0.7730047       -0.21588252
## Arunachal Pradesh          0.1255788      0.1772946       -0.40651414
## Assam                      0.8869829      1.6240190       -0.45868599
## Bihar                      1.9101196      1.1985118       -0.47075536
## Chandigarh                -1.0879090     -0.5886182       -0.05370242
##                           RoadLength
## Andaman & Nicobar Islands -0.8596848
## Andhra Pradesh             0.3558963
## Arunachal Pradesh         -0.6954134
## Assam                      1.3649922
## Bihar                      0.5405426
## Chandigarh                -0.8488543
```

4.FIND EUCLIDEAN DIST FOR SAMPLE DATA

```
ind <- sample(1:nrow(df),15,replace = T)
df15 <- df[ind,]
df15_dist <- dist(df15)
```

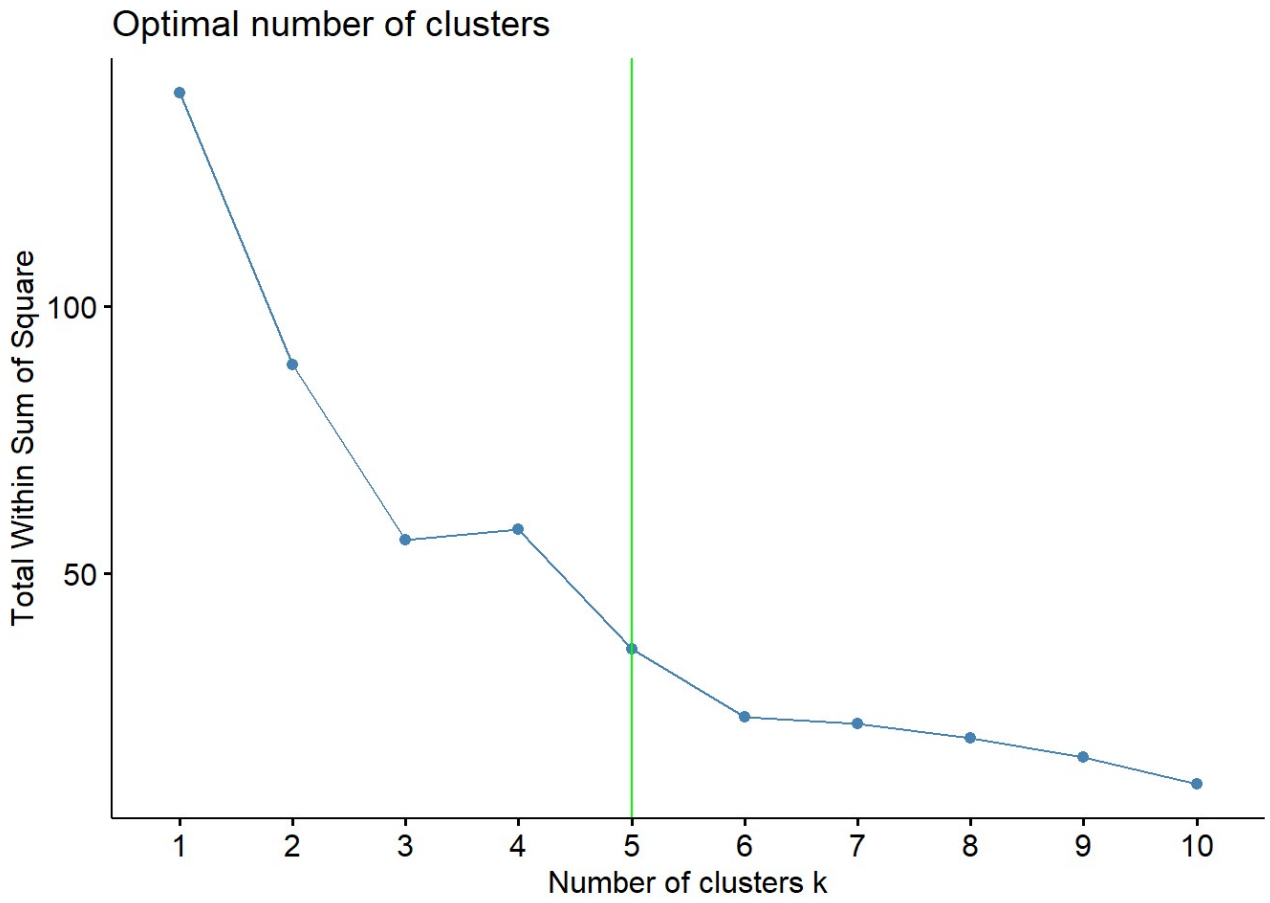
5.CHECK FOR CLUSTER FORMATION

```
fviz_dist(df15_dist)
```



6.NUMBER OF CLUSTER USING WSS

```
fviz_nbclust(df,kmeans,method='wss')+geom_vline(xintercept = 5,linetype=7,col='green')
```



7.K MEANS

```
df.km=kmeans(df,5,nstart=20)
df.km
```

```
## K-means clustering with 5 clusters of sizes 2, 10, 4, 14, 6
##
## Cluster means:
##    BirthRate MortalitityRate PowerAvailability RoadLength
## 1  0.7204257      -0.7162704         3.7851447 -0.8641901
## 2  0.3896908       0.6113119        -0.2248171 -0.3492352
## 3 -0.4871136      -0.3758646        -0.2133319  1.6745781
## 4 -0.8346741      -0.8743159        -0.1649130 -0.5974956
## 5  1.3826887       1.5105504        -0.3600015  1.1478930
##
## Clustering vector:
## Andaman & Nicobar Islands            Andhra Pradesh         Arunachal Pradesh
##                         4                         2                         2
##                     Assam                     Bihar               Chandigarh
##                         5                         5                         4
##              Chhattisgarh       Dadra& Nagar Haveli               Daman & Diu
##                         2                         1                         1
##                     Delhi                       Goa                   Gujarat
##                         4                         4                         2
##                   Haryana          Himachal Pradesh         Jammu and Kashmir
##                         2                         4                         4
##                 Jharkhand                 Karnataka                    Kerala
##                         2                         3                         4
##               Lakshadweep            Madhya Pradesh               Maharashtra
##                         4                         5                         3
##                   Manipur                 Meghalaya                   Mizoram
##                         4                         2                         2
##                  Nagaland                    Odisha                Puducherry
##                         4                         5                         4
##                    Punjab                 Rajasthan                    Sikkim
##                         4                         5                         4
##                 Tamil Nadu                 Telangana                   Tripura
##                         3                         2                         4
##             Uttar Pradesh               Uttarakhand               West Bengal
##                         5                         2                         3
##
## Within cluster sum of squares by cluster:
## [1] 5.069544 7.120893 4.249168 7.815848 3.958959
##  (between_SS / total_SS =  79.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
df.km$cluster
```

```
## Andaman & Nicobar Islands            Andhra Pradesh         Arunachal Pradesh
##                          4                         2                         2
##                      Assam                     Bihar               Chandigarh
##                          5                         5                         4
##              Chhattisgarh       Dadra& Nagar Haveli             Daman & Diu
##                          2                         1                         1
##                      Delhi                       Goa                   Gujarat
##                          4                         4                         2
##                    Haryana           Himachal Pradesh         Jammu and Kashmir
##                          2                         4                         4
##                  Jharkhand                 Karnataka                    Kerala
##                          2                         3                         4
##                Lakshadweep             Madhya Pradesh               Maharashtra
##                          4                         5                         3
##                    Manipur                 Meghalaya                   Mizoram
##                          4                         2                         2
##                   Nagaland                    Odisha                Puducherry
##                          4                         5                         4
##                     Punjab                 Rajasthan                    Sikkim
##                          4                         5                         4
##                  Tamil Nadu                 Telangana                   Tripura
##                          3                         2                         4
##              Uttar Pradesh               Uttarakhand               West Bengal
##                          5                         2                         3
```

8.ADD MEMBERSHIP TO THE DATA

```
RBI_1 <- cbind(RBI,Clusters=df.km$cluster)
head(RBI_1,3)
```

	States_Union <fctr>	BirthRate <dbl>	Mortalityl
Andaman & Nicobar Islands	Andaman & Nicobar Islands	12.0	
Andhra Pradesh	Andhra Pradesh	16.8	
Arunachal Pradesh	Arunachal Pradesh	18.8	

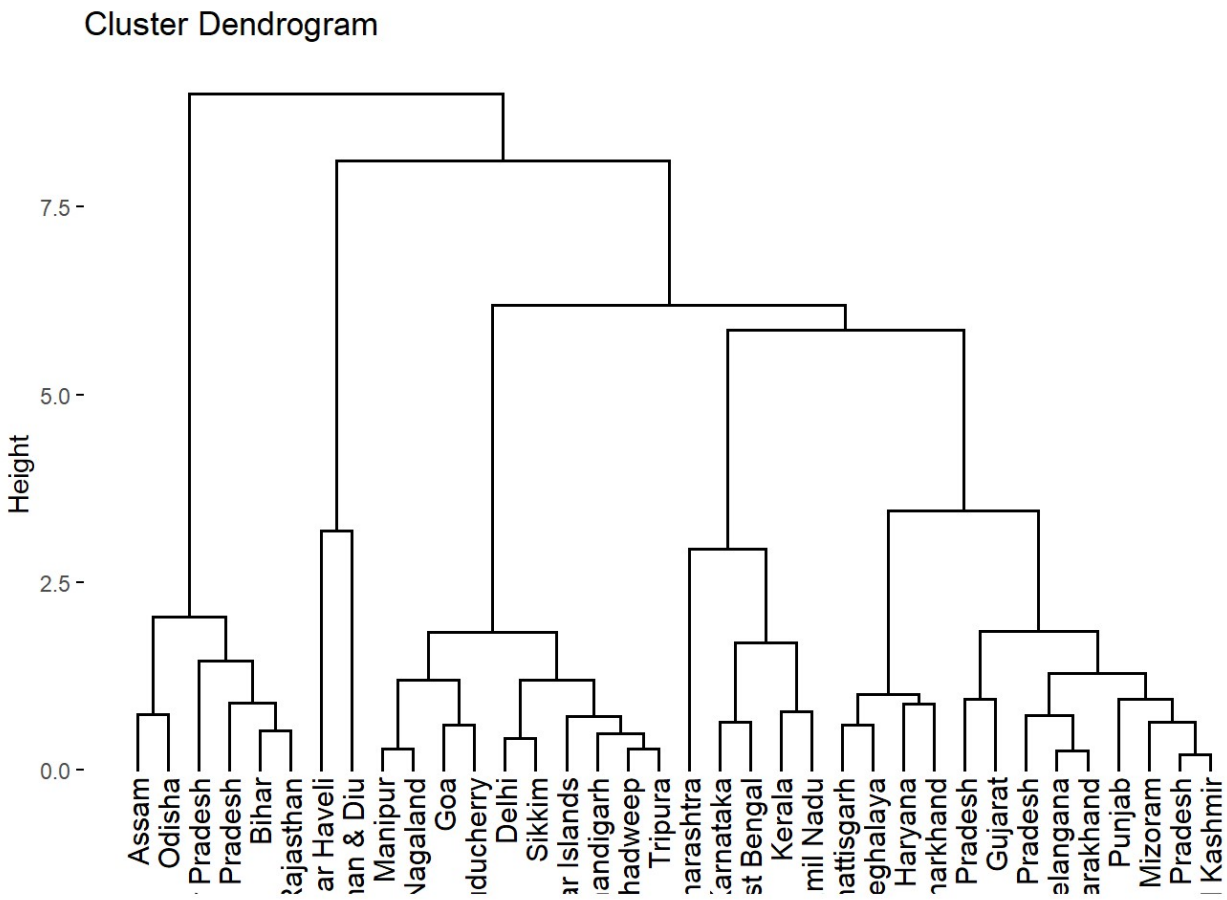
3 rows | 1-5 of 7 columns

9.HEIRARCHICAL CLUSTEING

```
df.dist <- dist(df,method="euclidean")
df.hc <- hclust(df.dist,method='ward.D2')
```

10.VISUALIZE DENDOGRAM

```
fviz_dend(df.hc)
```



11.CLUSTERING USING CUTREE

```
grp <- cutree(df.hc,k=5)
```

12.ADD MEMBERSHIP TO THE DATA

```
RBI_2 <- cbind(RBI,clusters=grp)
head(RBI_2,3)
```

	States_Union <fctr>	BirthRate <dbl>	Mortalityl
Andaman & Nicobar Islands	Andaman & Nicobar Islands	12.0	
Andhra Pradesh	Andhra Pradesh	16.8	
Arunachal Pradesh	Arunachal Pradesh	18.8	

3 rows | 1-5 of 7 columns

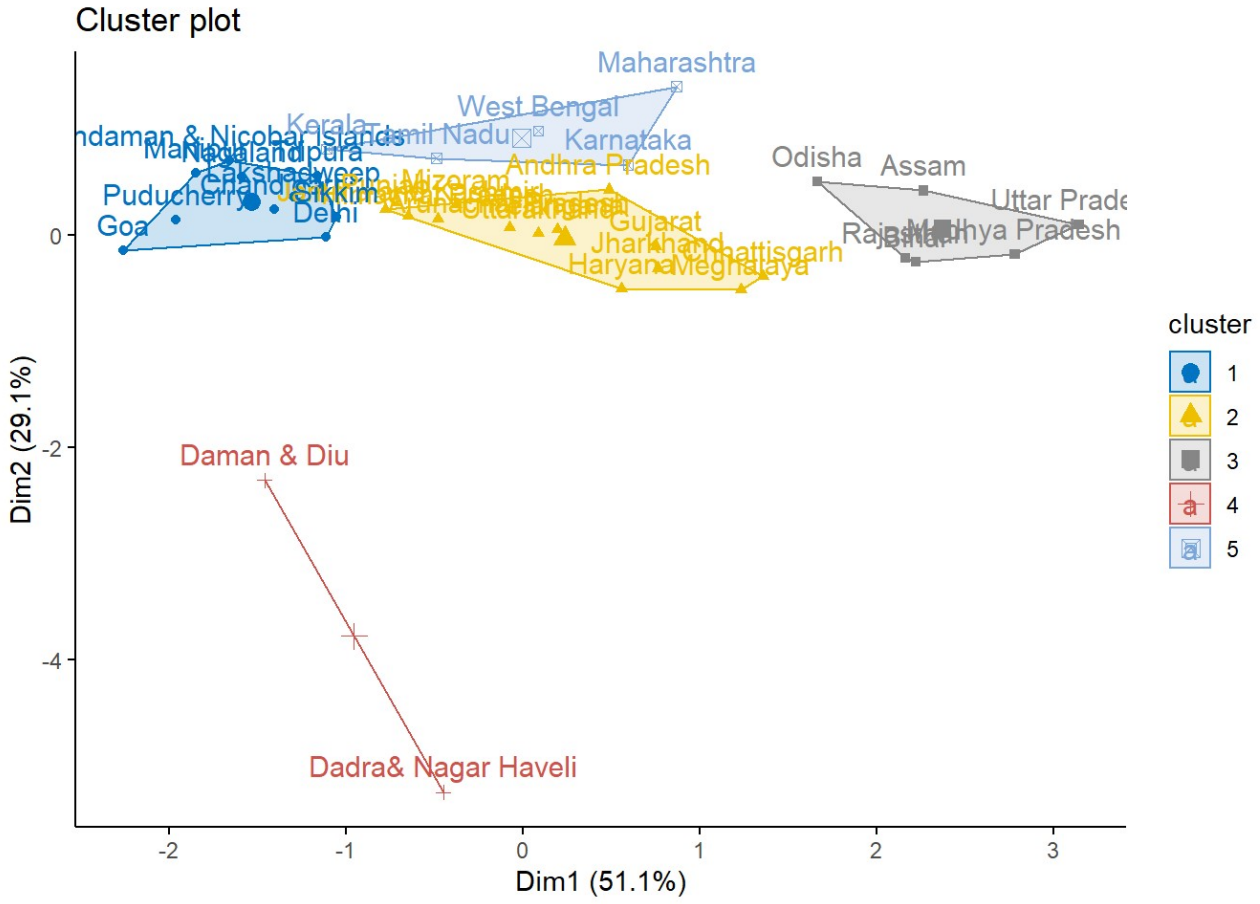
13.COMPARISION OF KMEANS AND HCLUST

```
table(df.km$cluster,grp)
```

```
##    grp
##      1  2  3  4  5
##   1  0  0  0  2  0
##   2  0 10  0  0  0
##   3  0  0  0  0  4
##   4 10  3  0  0  1
##   5  0  0  6  0  0
```

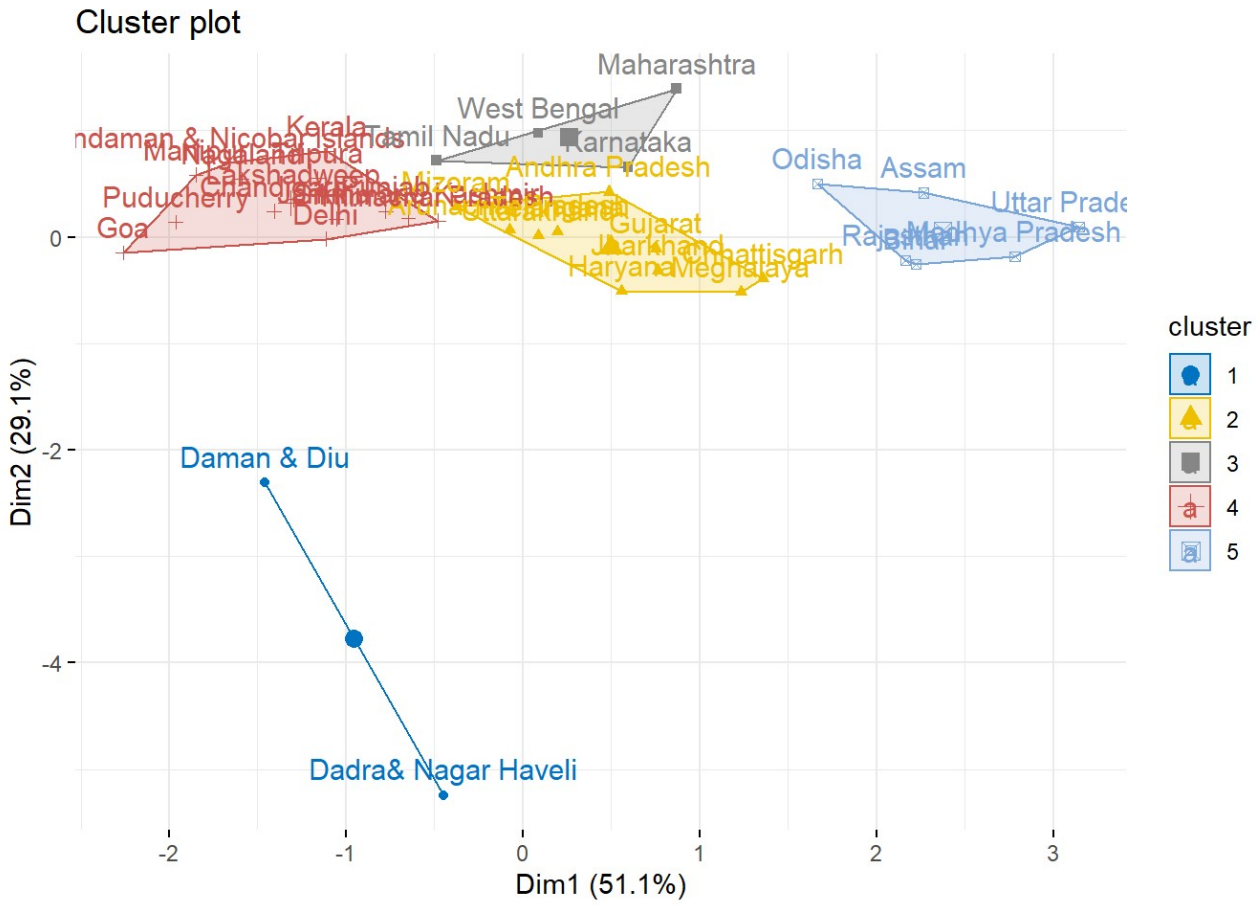
14.K MEANS CLUSTERING

```
fviz_cluster(list(cluster=grp, data = df), palette = "jco",ggtheme = theme_classic())
```



15.HIERARCHICAL CLUSTERING

```
fviz_cluster(df.km, data = df, palette = "jco",ggtheme = theme_minimal())
```



16.PROFILING KMEANS AND HCLUST GROUPS OF ORIGINAL DATA

```
aggregate(RBI[,-1],by=list(df.km$cluster),mean)
```

Group.1 <int>	BirthRate <dbl>	MortalityRate <dbl>	PowerAvailability <dbl>	RoadLength <dbl>
1	21.30000	19.50000	13451.1500	693.50
2	19.91000	35.10000	992.0400	75959.60
3	16.22500	23.50000	1027.7250	371761.25
4	14.76429	17.64286	1178.1643	39673.71
5	24.08333	45.66667	572.0167	294780.67

5 rows

```
aggregate(RBI[,-1],by=list(grp),mean)
```

Group.1 <int>	BirthRate <dbl>	MortalityRate <dbl>	PowerAvailability <dbl>	RoadLength <dbl>

Group.1 <int>	BirthRate <dbl>	MortalityRate <dbl>	PowerAvailability <dbl>	RoadLength <dbl>
1	14.42000	15.80000	1161.0700	16052.10
2	18.98462	32.92308	1085.3308	73819.46
3	24.08333	45.66667	572.0167	294780.67
4	21.30000	19.50000	13451.1500	693.50
5	15.94000	21.20000	961.1200	336379.80

5 rows

CONCLUSION 1.Kmeans Clustering has 5 clusters of sizes 10,13,6,2,5.

2.Heirarchical clustering 5 clusters of sizes 10,14,6,2,4.